1-1-2013

Collaborative creation of spoken language corpora

Michael Haugh
Griffith University

Wei-Lin Melody Chang
Griffith University, wlchang@uow.edu.au

Collaborative creation of spoken language corpora

Abstract

Analysing authentic interactions at progressively greater levels of complexity is one means of promoting deeper engagement with pragmatic phenomena amongst L2 learners. However, effective analysis often requires a greater amount of data than learners can feasibly gather. It is proposed here that encouraging students to collaborate through the creation of a corpus of spoken interactions is one potentially effective way to help them engage with a much richer set of interactional data than they might normally encounter. Here we report on a corpus created through "crowdsourcing" the collection and transcription of recordings of spoken interactions, the Griffith Corpus of Spoken Australian English (GCSAusE), which was then made available to L1 and L2 students to use in analysing pragmatic aspects of spoken interaction. In this way, the students had the opportunity to be both creators and users of the corpus, and see how it results in the real and ongoing accumulation of knowledge about language use. The degree of engagement of students with the corpus was assessed through their research projects, a written survey, and a focus group conducted with a number of students who took the course.

Keywords

language, spoken, creation, corpora, collaborative

Disciplines

Arts and Humanities | Law

Publication Details

Haugh, M. & Chang, W.M. (2013). Collaborative creation of spoken language corpora. In T. Greer, D. Tatsuki & C. Roever (Eds.), Pragmatics and Language Learning (Volume 13) (pp. 133-159). University of Hawai'i at Mānoa: National Foreign Language Resource Center.

Collaborative creation of spoken language corpora

Michael Haugh and Wei-Lin Melody Chang
Griffith University

Abstract:

Analysing authentic interactions at progressively greater levels of complexity is one means of promoting deeper engagement with pragmatic phenomena amongst L2 learners. However, effective analysis often requires a greater amount of data than learners can feasibly gather. It is proposed here that encouraging students to collaborate through the creation of a corpus of spoken interactions is one potentially effective way to help them engage with a much richer set of interactional data than they might normally encounter. Here we report on a corpus created through "crowdsourcing" the collection and transcription of recordings of spoken interactions, the Griffith Corpus of Spoken Australian English (GCSAusE), which was then made available to L1 and L2 students to use in analysing pragmatic aspects of spoken interaction. In this way, the students had the opportunity to be both creators and users of the corpus, and see how it results in the real and ongoing accumulation of knowledge about language use. The degree of engagement of students with the corpus was assessed through their research projects, a written survey, and a focus group conducted with a number of students who took the course.

1. Introduction

Research indicates that awareness of L2 pragmatic norms is not acquired through simply being immersed in an L2 environment, but requires sustained attention and effort from students to learn (Kasper & Rose, 2002). In this respect, it has been found that the development of pragmatic competence can be facilitated by explicit instruction, where learners are not only exposed to contextualised input, but are also encouraged to engage in (meta)pragmatic analysis of relevant phenomena (Ishihara, 2010; Ishihara and Cohen, 2010; Kasper, 2001; Rose, 2005). While there is some controversy as to which particular teaching approaches are more effective (Jeon & Kaya, 2006; Rose & Ng, 2001; Takimoto, 2008), having students analyse authentic interactions in their L2 at progressively greater levels of complexity appears to be one effective means of promoting deeper engagement with pragmatic phenomena. However, effective analysis of pragmatic aspects of interaction presupposes not only access to authentic interactions, but a greater amount of data than learners could feasibly gather themselves. It is thus proposed here that encouraging students to collaborate through the creation of a relatively large collection - or what is often termed a corpus - of spoken interactions is one potentially effective way of helping them engage with a much larger and more detailed set of interactional data than they might normally encounter. A corpus is generally defined as a relatively structured or targeted collection of samples of spoken (or written) data that is machine-readable, which means that large amounts of spoken interaction or texts can be searched according to specified parameters (Peters, 2009, pp. 1-2). Alongside the data itself, corpora also generally contain "meta-data", namely, information about the participants, time and place of recording and so on. Here we report on the development of a corpus, the Griffith Corpus of Spoken Australian English (GCSAusE), which was collaboratively created by students in order to collect and

transcribe recordings of naturally occurring spoken interactions (a process termed "crowdsourcing", Howe, 2006)

We begin by first outlining the case that has been made for drawing from authentic interactions in instructional pragmatics. The potential synergy between those who advocate applying results from studies in conversation analysis (CA), and those who have applied results from studies in corpus linguistics (or more specifically, corpus pragmatics), is also explored. One drawback of both approaches that emerges from this discussion is how to readily gain access to relevant datasets of authentic interactions. We next propose that recent developments in corpus building may offer at least one solution to this impediment to realising the promise of instructional pragmatics. After introducing the principles of cyclical and collaborative corpus creation, which allow for the progressive building of a corpus by multiple contributors, we next outline the implementation of these principles in the development of the GCSAusE, and discuss some of the practical problems that emerged in the course of building this corpus. The use of the corpus by advanced L2 and L1 students in a third-year university course in English pragmatics is then examined through multiple evaluative perspectives including: (1) analysis of research projects they conducted based on data from the corpus, (2) the results of a survey conducted with all the students in the course, and (3) a focus group conducted with a mixture of L1 and L2 speakers from that course. We conclude with a brief discussion of the promise that such an approach offers for instructional pragmatics more broadly, along with some of its limitations.

2. Authentic interactions in the teaching of L2 pragmatics

The view that we should be emphasizing the use of authentic interactions in teaching L2 pragmatics is now generally advocated by many if not most applied linguists. There are, however, differences in what is considered authentic and/or interaction. The teaching of L2 pragmatics has traditionally used data generated through discourse completion tests (DCTs), or alternatively examples constructed through native speaker intuition. One key problem with using such data, however, is that it "affords somewhat idealized versions of social interaction" (Huth & Taleghani-Nikazam, 2006, p. 54). Such idealised sociopragmatic norms may be incongruent with what actually occurs in interaction, as amply demonstrated by both conversation analysts (Kasper, 2006; Huth & Taleghani-Nikazam, 2006), and those working in corpus pragmatics (Adolphs 2008; Geluykens & Kraft, 2008; Vine 2004, 2009; cf. Schauer & Adolphs, 2006), for instance.

A second problem is that such data isolate the analysis of pragmatic phenomena from their sequential environment. As Kasper (2006) argues, traditional approaches based on speech act theory do not fully account for the "indexical character of situated action and especially its sequential environment" (p. 297). This is because speech acts are often not accomplished through a single turn at talk, but can be co-constructed over multiple turns (Huth & Taleghani-Nikazam, 2006, p. 63). Moreover, the temporal structure of actions in turns can also be critical to the analysis of various pragmatic phenomena (Kasper, 2006, p. 297). Such phenomena include meaning beyond what is said (what is presupposed, implied, or referred to through what is said), social actions (both those that form a part of members' conscious metapragmatic awareness and those are a part of only their interactional competence), and the evaluation of persons and relationships in conversational interaction (encompassing im/politeness, facework, humour, relational identity and the like)

(Haugh, 2012).[1] In taking the position that meanings, actions and evaluations are interactionally and situationally achieved, that is, they are "constituted not only *in* but *through* social interaction" (Kasper, 2006, p. 282; see Arundale, 1999, 2005, 2010 for further discussion), it is clear that in order to fully understand pragmatic phenomena we need to be drawing from authentic interactional data situated in their sequential context.

Nevertheless, no matter what stance one ultimately takes on the issue of what counts as authentic interaction, there are numerous challenges facing any teacher wanting to introduce such data into the classroom. One key problem that emerges from an examination of studies that have strongly advocated the use of authentic interactions is just how one can obtain sufficient data for use in the classroom. Here we focus our discussion on challenges facing those advocating the use of conversation analysis (Barraja-Rohan, 1997; Félix-Brasdefer, 2006; Huth & Taleghani-Nikazam, 2006; Kasper, 2006; Koester, 2009; Wong & Waring, 2010), or corpus pragmatics (Geluykens & Kraft, 2008; Holmes, 2009; Jiang, 2006; Koester, 2002; Marra, 2008; Newton, 2004; Usami, 2005) in teaching L2 pragmatics, as these two disciplines are most directly relevant to the approach to corpus creation we advocate here.

A number of studies have illustrated the potential for detailed analyses of authentic interactions in the CA tradition to contribute to the teaching of L2 pragmatics. Huth and Talgehani-Nikazm (2006), for instance, argue that guiding learners through contrastive analyses of the opening sequences of telephone calls amongst speakers of American English and those between Germans, not only provides them with a "blueprint" for this particular conversational action sequence. CA-based analyses enable learners to identify key differences that do not figure in standard textbook accounts (Wong, 2002), as well as raising awareness amongst language teachers about the folk or pre-scientific understandings of pragmatic phenomena that dominate textbooks (Yates, 2010, p. 129). There are, however, some natural limitations to this approach, at least as framed by scholars thus far. Firstly, the way in which pragmatic phenomena are selected for teaching seems somewhat opportunistic, in the sense that the teacher here is actually directly drawing from his/her role as researcher. Félix-Brasdefer (2006), for example, draws from his broader (2008) study of refusals in American English and Mexican Spanish, while Huth and Talgehani-Nikazm (2006) draw from their earlier research on telephone openings in German (Talgehani-Nikazm, 2002). Secondly, the range of pragmatic phenomena selected for teaching seems to be limited in the literature thus far to the core concerns of CA, namely, turn-taking, adjacency, topic management, story-telling, openings and closings, repair, and a limited number of social actions (see, for instance, Wong & Waring, 2010). While such concerns are indeed important, pragmatic competence encompasses a broader range of phenomena that lie outside the direct purview of CA, in particular, meanings beyond what is said (implicature, reference, deixis, presupposition), as well as interpersonal evaluations (im/politeness, face practices etc.). An over-reliance on CA-based materials thus potentially limits the scope of what is taught.[2] Thirdly, CA datasets are for the most part closed, in the sense that while transcripts are available for inspection, the original recordings are generally not made available beyond select groups of researchers, let alone to groups of L2 learners. This is not meant as a criticism of CA research per se, as researchers often have legitimate reasons why audio (visual) data cannot be made be widely available, often relating to the conditions imposed by the participants in the recordings. An important exception to this trend is Schegloff's website, where he makes audio/visual files, which feature mainly speakers of American English,

available alongside his published work, as well as Talkbank, which provides access to an increasing number of CA transcriptions and audiofiles.[3] However, at present learners do not generally have the opportunity to listen to accompanying recordings, except when the teacher is also the researcher in question. The problem is that exposing learners to CA transcripts without the opportunity to listen to original recordings not only makes it difficult for those learners to interpret and understand those interactions in the first place, it is also inconsistent with the insistence of CA practitioners that the data for analysis ultimately reside in the recording not the transcript.

A parallel move towards using authentic interactions in the teaching of L2 pragmatics has emerged in applications of methodologies in corpus linguistics to pragmatics, or what has recently come to be known as corpus pragmatics (Jucker, Schreier & Hundt, 2009; Romero-Trillo, 2008; Rühlemann, 2010). Pedogogical applications of corpus pragmatics have advocated going beyond teaching lists of speech act phrases or syntactic structures to considering the relative frequency of different syntactic structures as well as illocutionary force in different registers and genres (Jiang, 2006, Koester, 2002; Usami, 2005). The importance of exposing learners to recordings and transcriptions of authentic interactions is also argued by those advocating corpus-based approaches to teaching pragmatics. A corpus-based approach also allows for the examination of frequency of particular collocations in different contexts as well as common sequential structures underlying speech acts. Such an approach can thus be used to assist in the identification (Wulff, 2010), and teaching of formulaic or conventional expressions (Chambers, 2007; O'Keefe, McCarthy & Carter, 2007; Schmitt, 2004), the use of which is often avoided by L2 learners (Bardovi-Harlig, 2006, 2009, 2010, this volume).

However, while such studies demonstrate the great potential for corpus linguistics to contribute to the teaching of L2 pragmatics, there are also some natural limitations in the studies published thus far, particularly in regards to spoken interaction. Firstly, transcripts in spoken corpora generally lack sufficiently detailed paralinguistic, nonverbal and contextual information (Geluykens & Kraft, 2008). This means more detailed analyses of the interactional accomplishment of pragmatic actions is difficult (Adolphs & Carter, 2007; Haugh, 2009). In some spoken corpora certain features of spoken interaction, such as overlap, pauses, or laughter, are included in transcripts, such as in the case of the spoken components of the British National Corpus or International Corpus of English (Crowdy, 1993, 1994). However, the level of detail in the transcriptions is minimal. In other corpora, such details are completely absent. The Corpus of Contemporary of American English, for example, which features 85 million words of spoken text, draws from transcripts of what is said only (Davies, 2009). A related problem is that the primary focus of analysis in spoken corpora has always been textual transcriptions of audio (visual) recordings, with the original recordings themselves not traditionally being considered the focus of analysis, and so are generally not made widely available (Wichmann, 2008, p.189). The only exception to this, at least in relation to English, is the Santa Barbara Corpus of Spoken American English, where audio files are made available alongside more detailed transcriptions.[4] Spoken corpora are thus almost always closed in the sense that audio recordings are not accessible, except to select researchers. Thirdly, the lack of direct correspondence between linguistic forms and pragmatic phenomena means the applications of linguistic corpora to instructional pragmatics are still somewhat limited (Ishihara, 2010), with phraseology, backchannels, and discourse markers receiving the most analytical attention in corpus pragmatics thus far (Rühlemann,

2010). While there have been recent attempts to study other pragmatic phenomena using corpora, including speech acts (Adolphs, 2008; Ramírez-Verdugo, 2008; Schauer & Adolphs, 2006; Vine, 2004), humour (Vaughan, 2008), and im/politeness (Clancy, 2011; Culpeper, 2011; Taylor, 2009, 2011), all of these studies draw from spoken corpora that are not accessible to teachers or learners. The corpora are either private collections of the researchers themselves, or are closed corpora (i.e. not made available beyond select groups of researchers), such as the Cambridge and Nottingham Corpus of Discourse in English (CANCODE) (Schauer & Adolphs, 2006), or the Language in the Workplace Corpus (Newton, 2004).

In summary, then, both CA-based and corpus-based approaches to instructional pragmatics advocate using authentic interactional data, albeit using quite different methods of representing and analysing such data. However, their call to arms is clearly attenuated by the simple fact that such materials, in particular, original audio (visual) recordings, are not actually widely available, either to language teachers or to L2 learners. This problem is, of course, not limited to those advocating CA-based or corpus-based approaches to teaching pragmatics, but is a challenge facing all those who advocate communicative language teaching more broadly. It appears, then, that the very real promise of using authentic interactional data in teaching L2 pragmatics is being hampered by the issue of where to source such materials.

In the following section, we suggest that cyclical and collaborative corpus creation, where the learners themselves are involved in the process of gathering and analysing spoken interactions, offers one potential solution to this data bottleneck in instructional pragmatics, particularly for more advanced L2 learners. It is also suggested that such a process enables language teachers to build on the relative strengths of CA in regards to the level of detail in transcription, on the one hand, and corpus pragmatics, which enables targeted search across relatively large datasets, on the other.

3. Cyclical and collaborative corpus creation

Spoken corpora have traditionally been very time-consuming and expensive to build. While building a representative collection of spoken interactions in the order of the ten million word spoken component of the British National Corpus is clearly not called for in teaching L2 pragmatics, a certain minimal amount of data is nevertheless required for analysing pragmatic phenomena. The problem is that even constructing a fairly limited or specialised corpus of spoken interaction involves collecting a greater amount of data than individual teachers or learners could feasibly gather on their own. Yet, as most of the corpora of spoken interaction that have been created to date are not readily accessible, it appears that collecting one's own data remains necessary if we are to realise the promise of an instructional pragmatics that is grounded in authentic interactional data.

One alternative to the traditional approach to building spoken corpora, however, is to employ a cyclical and collaborative model of creation. Instead of attempting to create a complete spoken corpus in its entirety before it can be used (i.e., traditional sequential corpus creation), a cyclical process model is proposed as a more realistic model in the context of teaching L2 pragmatics. This collaborative and cyclical model involves two key stages:

Stage 1: Crowdsourcing the recording and transcription of spoken interaction.

Stage 2: Query-driven, progressive annotation of relevant pragmatic features in those spoken interactions.

The first stage involves asking multiple contributors to record and transcribe one or two interactions each (i.e. crowdsourcing), thereby building a large collection through strength in numbers. The second stage refers to a bottom-up approach to adding pragmatic information (i.e. annotations) about the interactions based on the interests of the students themselves (i.e. query-driven) rather than being top-down or theory-driven. This model draws from Brinckmann's (2009) crowdsourcing model of transcription, and Voormann and Gut's (2009) Agile Corpus Creation Theory, which are discussed in further detail in sections 3.1 and 3.2 respectively.

It is worth reiterating at this stage that the term corpus can actually be used to refer either to a "structured collection of texts sampled from various types of [spoken] discourse" (Peters, 2009, p. 1), which "aim[s] for as broad, balanced and comprehensive coverage of spoken language data as possible that can later be used for many types of balanced and representative research" (Čermák, 2009, p. 114), or a largely ad hoc assemblage of spoken texts often associated with a particular research project (Peters, 2009, p. 1). The model proposed here begins by creating a corpus in the second more ad hoc sense, with the view to eventually creating a corpus in the first more structured and representative sense.

It is also worth noting that this model assumes a distinction between transcriptions and annotations, both of which are made in order to allow for the search and analysis of pragmatic phenomena in audio (visual) recordings. Transcriptions, which are fundamental to CA, and have been commonly used in building traditional spoken corpora, involve the representation of speech in textual form, including what is said, syntactic and lexical units, prosodic features, as well as (sometimes) nonverbal aspects of interaction. Annotations, on the other hand, are machine-readable, text-based pointers to such features in the audio (visual) files, themselves. They can be used to identify a broader range of pragmatic phenomena than transcriptions, however, as they also include descriptors of longer sequences, such as speech acts or activity types, for instance. However, regardless of whether one chooses to create transcriptions or annotations, the same issue arises, namely, that both transcribing and annotating spoken interaction is largely a manual and time-consuming process (Allwood, 2008; Brinckmann, 2009; Thompson, 2004). The cyclical and collaborative model of spoken corpus creation is proposed here as a way of sharing the load as it were.

3.1. Crowdsourcing recording and transcription

Crowdsourcing refers to outsourcing a task to a large, sometimes undefined, group of people. It may also take advantage of Web 2.0 technologies. In regards to the creation of spoken corpora, both the gathering of recordings and their transcription can be crowdsourced.

Brinckmann (2009) makes reference to the PHATT speech database of German teenage speech (collected primarily for phonetic analysis), where examples

of read and spontaneous speech were recorded by participants on their PCs with an Internet connection, as an example of how this principle can be put into practice. The recordings themselves were prompted via a web-based speech recorder (now part of WikiSpeech: http://wikispeech.org), which uploaded the recordings to a central server. In this way, the participants themselves were able to carry out the recordings without a researcher or technician being present (Brinckmann, 2009, p. 68). There is, however, another potentially less technologically-constrained way of crowdsourcing the collection of spoken recordings, which draws from a source readily available to language teachers, namely, students. In this case, one can ask a group of students to go out and make such recordings as part of their coursework. It is the latter approach which was utilised in the creation of the Griffith Corpus of Spoken Australian English (GCSAusE).

The transcription process can also be crowdsourced. Brinckmann (2009) cites the example of the "German Today" speech project where a system was set up to enable efficient crowdsourcing of transcription. This system is represented in Figure 1 below.

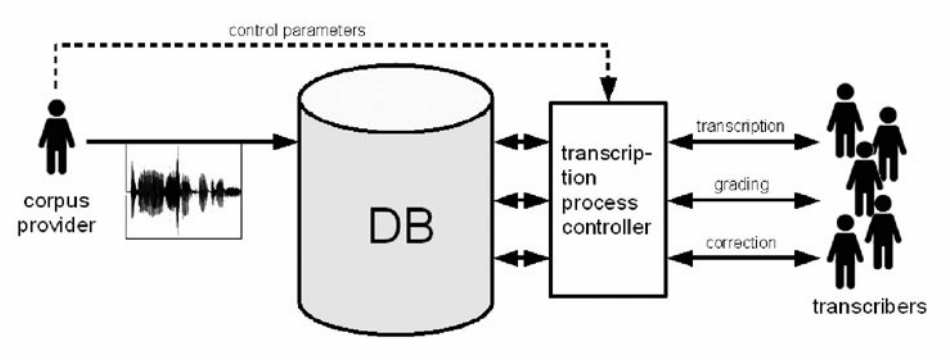

Figure 1: Simplified architecture for system for crowdsourcing transcription (Brinckmann, 2009, p. 172)

The above system consists of six key components:

a. A database of speech files and metadata.
b. Task definition (including conventions for transcription, grading and correcting tasks).
c. Process control.
d. Database of human transcribers.
e. Transcription process (initial transcription, grading, correction).
f. Rewards (grades, lists of top transcribers etc.) (Brinckmann, 2009, pp. 170-1)

In this system, contributors may transcribe recordings they themselves have made, or others provided by the teacher. The system of transcription can vary, of course, with standard CA transcription conventions (Jefferson, 2004) being more suited to advanced level learners, while less advanced learners might use a more simplified transcription system. There are three modes of presentation of transcriptions (or annotations): the "vertical mode" is generally employed by those who use text-processing software like Word to create transcriptions, while the "partiture mode" (which is similar to an orchestral score) or the "column mode" are utilised in specialised annotation editors, such as EXMARaLDA, ELAN, CLAN or Transana

(Rohlfing et al., 2006). One problem with the fairly widespread practice of creating vertical transcriptions in programs like Word is that such documents are not readily machine-readable, nor are they easily time-aligned with the original recordings (Haugh, 2009, p. 80). The use of specialised annotation editors, on the other hand, allows the researcher to create machine-readable annotations with varying degrees of interoperability across different software systems. However, the use of specialised annotation software may not be feasible in language classrooms, as it can take some time to learn how to use such software. Moreover, the software and thus the annotations it generates may become outdated or no longer supported (Deppermann & Schütte, 2008, p. 198). The creation of traditional vertical CA transcriptions in Word thus remains a practical compromise for the moment, although it does limit the addition of further pragmatic information, such as speech act descriptors and the like.

After a transcription is completed and submitted, it can be graded by the teacher (or alternatively peer-reviewed) and then sent back to the learner for correction. This process of correction can also be useful from a pedagogical perspective as it allows the teacher to draw the learner's attention to features that he/she did not notice in the first instance, or to their "mishearings" of certain parts of the recording.

In order for the crowdsourcing of the recording and transcription of spoken interaction to be successful, Brinckmann (2009) suggests that three general principles be followed:

1. Focus: Every task should be described as clearly as possible together with a set of rules.
2. Filter: Use the crowd and experts to extract the best answers.
3. Reward: can be money, recognition or fun. (pp. 169-170)

In the case of language classrooms, these three principles can be realised as follows: (1) the focus of the task should be carefully outlined, with prior training of the learners or students before they go out to record and transcribe the spoken interactions; (2) careful filtering of the recordings and transcriptions needs to be undertaken by the teacher, although peer-review of transcriptions is another pedagogical a practical means of sharing the load to ensure accurate transcriptions enter the corpus; and (3) the reward can be graded assessment, as well as the satisfaction gained from contributing to something both oneself and others are able to later collectively use.

Once an accurate transcription of the recording is available it can be added to the corpus together with the original recording itself, as well as metadata about the recording (e.g., when and where the recording took place, the background of the participants etc.). At this point the corpus can be searched (or queried), and excerpts identified by both the teacher and learners. The ultimate use these excerpts are put is dependent, of course, upon the pedagogical model employed in the classroom.

In order to add further value to the collection, however, further annotation is necessary if learners are to readily identify examples of pragmatic phenomena in the corpus. In the cyclical model proposed here it is suggested that such pragmatic information can be progressively added in the form of annotations based on the interests of the learners themselves.

3.2. Query-driven progressive annotation

Traditionally in corpus creation, an annotation schema (i.e., a structured set of inter-related categories applied to different pragmatic phenomena) is first created and then applied to a set of data. In Agile Corpus Creation Theory (Voorman & Gut, 2008), however, it is proposed that the sequential process be replaced by a cyclic process model that is driven by queries from users of the corpus, as illustrated in Figure 2 below.

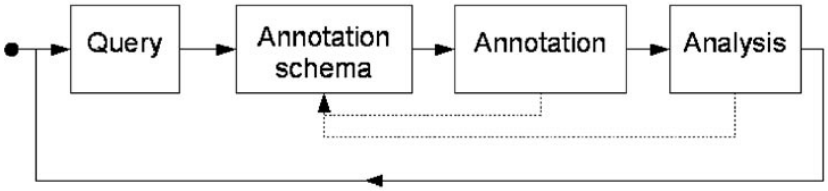

Figure 2: Query-driven corpus annotation process (Brinckmann, 2009, p. 175)

The annotation process starts with a query from the learner, which is then re-defined if necessary relative to a basic annotation schema, subsequently added as an annotation to the corpus, and finally fed back into the learner's analysis. Although allowing annotation to be progressively created in this way diverges from the traditional approach to corpus creation, a query-driven approach to corpus building is likely to yield a spoken corpus that can be put to work quickly, and also avoids inadvertently building early errors in the annotation process into the whole corpus (Voorman & Gut, 2008, p. 235). The aim of this approach is to allow "successive cycles [to] improve the annotation scheme and limit it to the elements necessary for the queries" (Brinckmann, 2009, p. 175).

While the employment of established standards is recognised as the ideal for most types of annotation (Haugh, 2009, p. 81), there is much less certainty in regards to adding pragmatic annotations for the simple reason that such annotations are essentially a kind of interpretative record, and thus are always embedded within a particular analytical and theoretical perspective (Archer, Culpeper & Davies, 2008, p. 637). This means a theory-driven pragmatic annotation system may inadvertently fail to identify important phenomena in the data. A query-driven approach to pragmatic annotation avoids this problem, at least to some extent, as it is driven through bottom-up analysis of the spoken data.

At a very minimum, pragmatic annotation can be used to create a record of social actions identified by learners in the spoken recordings. Some of these social actions are familiar to us through vernacular labels (e.g. requests), while others go beyond the metapragmatic awareness of ordinary members (e.g. reformulations) (Kasper, 2006, p. 305). Various pragmatic annotation schemas are available (Archer et al., 2008), which can offer a source of labels to consistently identify pragmatic phenomena in the corpus. A limitation of such annotation schemes, however, is they do not necessarily accommodate pragmatic phenomena that go beyond vernacular labels or particular theoretical models of language use.

In the following section, we outline how the Griffith Corpus of Spoken Australian English was created by implementing this cyclical and collaborative model, as well as briefly discussing some of the practical problems we faced in this process.

4. The Griffith Corpus of Spoken Australian English (GCSAusE)

The Griffith Corpus of Spoken Australian English (henceforth GCSAusE) is a progressively growing collection of audio recordings of face-to-face, interpersonal interactions between family members and friends conducted in homes and on university grounds in Brisbane, Queensland. The collection includes recordings made 2007-2010 inclusively. The participants are Australian speakers of English, although not all are Australian-born, reflecting the demographic reality of Australia, where up to 25% of Australians are born overseas. These audio recordings are accompanied by transcriptions made using standard CA conventions (Jefferson, 2004), along with metadata outlining basic information about the participants themselves, their relationships to each other (held in separate metadata records), as well as the locations and occasions of the conversations (listed at the beginning of the transcripts). The corpus is managed in an institutional repository system, the Equella-based Research Data Management System, which makes metadata about the GCSAusE publicly available, but not the transcripts or audio files. The entry portal for the GCSAusE is illustrated in Figure 3 below.

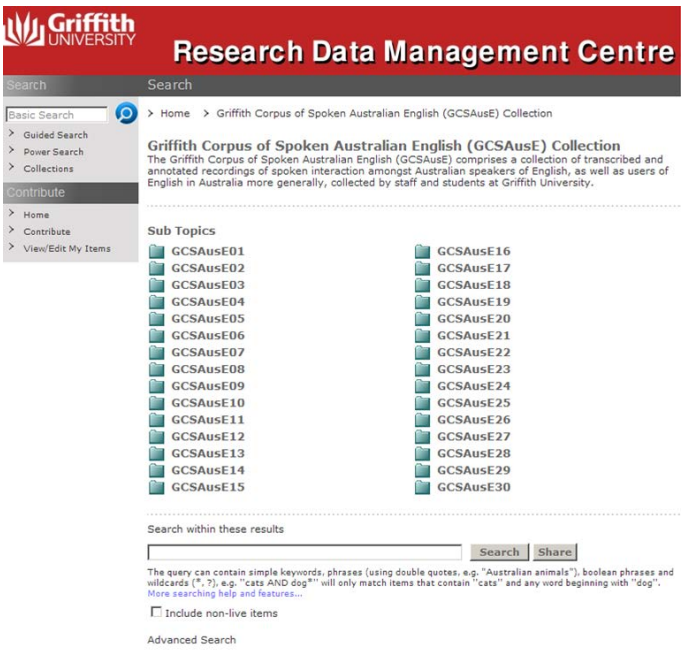

Figure 3: Screenshot of Griffith Corpus of Australian English entry portal

Access to audio recordings and transcripts is currently restricted to staff and students at Griffith University due to administrative limitations in setting up an appropriate access control system to the repository system. [5] The 2010 version of the GCSAusE consists of 30 recordings of spoken interactions, each of which is approximately five minutes in length, along with transcripts and associated metadata. The metadata for each recording consist of a separate entry about the audio file itself, the transcript, and each of the participants. Pseudonyms are used in the transcript and participant metadata records, while instances where the participants are explicitly named are muted in the audio recordings, in accordance with university ethics requirements. The corpus can be queried either through keyword search or by restricting searches

according to participant-related criteria (for example, a search restricted by gender or age).

The crowdsourcing of the recording and transcription production was accomplished in a third year course in English pragmatics, in which both L1 and advanced L2 speakers of English were enrolled. All students in the course are required to collect and transcribe spoken data as part of their assessed coursework. All the recordings and transcriptions were graded by the teacher (the first author) and necessary corrections indicated, with corrections also being suggested through peer review of the transcriptions. The contribution of their recordings and transcriptions to the GCSAusE was entirely voluntary, because it was dependent on gaining written consent from all the participants in the recordings themselves, as well as that of the student who made the recording in the first place. Roughly two thirds of the students offered to contribute their recordings and transcriptions in the 2009 and 2010 offerings of this course, although a small number of these were not accepted on the basis that transcriptions were not sufficiently accurate. Only audio recordings are included in the corpus. Although audiovisual recordings were sometimes made by students, inclusion of these in the GCSAusE was considered to be problematic, since ensuring anonymity for the participants is not possible in the case of audiovisual recordings. [6]

The requirement that transcriptions be consistent with standard CA conventions, and that the audio recordings be made available alongside the transcriptions means the GCSAusE differs from standard practices to date in building spoken corpora. In traditional spoken corpora, transcripts contain less detail about paralinguistic features of the interaction, for instance (cf. Crowdy, 1994), and audio recordings are also not readily accessible. The importance of having a more detailed transcript and accompanying audio files for the analysis of pragmatic phenomena, however, has been firmly established through a multitude of studies in CA and pragmatics (Haugh, 2009; Kasper, 2006), although this is perhaps not yet fully appreciated amongst corpus linguists (cf. Adolphs, 2008; Rühlemann, 2010).

In the following excerpt from an interaction between two Australian male housemates, taken from the GCSAusE, one calls the other a “nobhead”.

(1) GCSAusE06: 1:03

```
23    N:    so you were born
24          on Sunday, (0.5) of the fir:st month, (0.5) of (.)
25          the twenty-seventh day of nineteen eighty three=
26    D:    =↑no:, not ↑February ma:n
27          (0.2)
28    N:     oh, yo:u're a nobhea:d.
29          (0.6)
30    D:     °what° (.) h ha ↑hehehehe .hhhh
```

Up until this point in the conversation, Nick has been showing David how his new mobile phone can be used to calculate the day of the week on which David was born, which turns out to be a Sunday (lines 23-25). The insult in line 28 is occasioned by David’s slipup in thinking the first month of the year is February (line 26). David responds after a brief pause by delivering an open-class repair initiator (“what”, which orients to Nick’s insult in line 28), before displaying realization through his laughter that he has made a mistake. In this case, then, we have an instance of “jocular abuse”, that is, a non-serious insult (Haugh, 2009, pp. 77-8). However, without access to this kind of detailed transcript and accompanying audio recording such an analysis could only be tentative at best.

One advantage of including detailed CA transcripts in the GCSAusE is that they can be adapted to create more simplified transcripts quite readily, while the reverse is not the case. The disadvantage of the traditional CA approach to transcription, however, is that these details are not generated in the form of machine-readable annotations. The use of specialised transcription software, in particular, EXMARaLDA (Schmidt, 2009), was thus considered in a pilot study with a small number of users, but it was found to be very time-consuming to learn, and thus not appropriate in the context of a university course with limited contact hours.[7] The more traditional method of creating standard CA vertical transcriptions within Word documents was thus favoured despite its current limitations in regards to creating annotations. Students were, however, guided to use Audacity, a sound editor, to assist them in making their transcriptions.[8]

A number of practical problems were encountered in the course of crowdsourcing the creation of the GCSAusE. The first was that some students had difficulty producing completely accurate transcriptions. While one might expect that L1 speakers of English to be better placed to produce accurate transcriptions than L2 speakers, this was not in fact always the case. This is perhaps a reflection of their advanced level, as they were third year students taking the course in English pragmatics as part of an International English major specifically designed for L2 speakers. Nevertheless some of the L2 students did have difficulty producing completely accurate transcriptions. However, such difficulties occasioned opportunities for feedback from the teacher on aspects of spoken interaction they were unaware of or possibly mishearing, and so in that sense, what was a problem for building the corpus itself, represented an opportunity for learning on the part of students. A second issue was that many students had difficulty producing consistently formatted transcripts. The formatting of transcripts is important to ensure the accuracy of searches across data in the corpus. Some transcriptions were thus excluded from the GCSAusE for this reason, although peer review proved a useful means of improving the formatting of transcripts. A third problem was that students did not always provide sufficient details about the participants and the recording itself (i.e. the metadata), although this was more easily rectified by following up such details in class.

After its creation, the corpus was then made available to students to use in analysing pragmatic aspects of spoken interactions in Australian English in the 2010 offering of the course. At this point queries were created by students that are seeding the development of further pragmatic annotation, in particular, for different kinds of social actions found in the corpus. Pragmatic tags generated included practices such as “asking socially sensitive questions” (GCSAusE19, GCSAusE29), “broaching emotionally-charged topics” (GCSAusE19), “ironic receipting of complainables” (GCSAusE09), and “occasioning self-talk through inquiring about others” (GCSAusE09, GCSAusE15). However, while such tags contain useful pragmatic information, they are largely ad hoc, meaning systematic search across these annotations remains difficult for the simple reason students do not always know what they can be searching for across the corpus. The ad hoc nature of these pragmatic tags reflects the more general problem that we currently lack widely agreed upon standards for pragmatic annotation (Archer et al., 2008). In the following section we discuss a further potential issue, namely, the degree of engagement by students with the corpus, in particular, the L2 speakers.

5. Using the corpus: a student-based perspective

The GCSAusE was created by students enrolled in a third year course in English pragmatics. These students include both L1 and advanced L2 speakers of English. The former take the course as part of a major in linguistics, while the latter take it as part of a major in International English designed for L2 speakers. A key feature of this course is that it employs a research-based learning paradigm. In other words, students learn through conducting analyses of authentic interactional data themselves. In that sense, this approach is clearly most suited to educational contexts where there are advanced L2 speakers who specialise in English at undergraduate or even postgraduate level. The degree of engagement of students with the GCSAusE was thus considered to be fundamental for the relative success of the course from a pedagogical perspective. Their degree of engagement was evaluated in multiple ways, including through (1) an examination of the actual research projects they produced using the corpus, (2) a written survey which all the students taking the course answered, and (3) a focus group conducted with a small number of students in that course.

In the research projects, students were required to first record and transcribe a short, spoken interaction, and then identify a pragmatic phenomenon (e.g., a particular social action, instances of anticipatory completions, or a face practice) of interest in their own dataset. They were also required to find other examples of the same phenomenon in other data held in the GCSAusE for inclusion in their final analytical report. As noted previously, a number of practices were identified by students in their own data and in other recordings in the corpus. For example, two different practices were identified by students in the same excerpt from the corpus (GCSAusE09). The first was glossed "occasioning self-talk through inquiring about others" (initially noted by the student in data from the corpus, GCSAusE15), and the second "ironic receipting of complainables" (initially noted in the student's own data which was not subsequently contributed to the corpus).

The practice of occasioning self-talk through inquiring about others was first noticed by the student (an L1 speaker) in another conversation from the corpus (GCSAusE15), where two male housemates are chatting at home. The practice involves cases, as seen in the short excerpt below, where speaker A's inquiry about what speaker B has been doing (line 4) is not reciprocated by speaker B (line 5), yet speaker A nevertheless goes on to topicalise what he/she has been doing in a subsequent telling (line 6).

(2) GCSAusE15: 0:02

```
4      J:   been fishin' lately? (0.8)
5      N:   No: (0.6) I rea:lly wanna go fishin' actchally?
6      J:   been fishin' a fair bit down the coa:st
```

In example (3) taken from GCSAusE09, the student identified the same practice appearing over the course of a longer sequence. The relevant excerpt from a conversation between two male undergraduate students chatting at university is reproduced below.

(3) GCSAusE09: 0:00

```
1     B:    >so what did you do on the weekend<
2           (1.2)
3     A:    ah::: went and saw a friend and °ah:°
4           (0.6)
```

```
5     B:    ah ↑o↓kay, (1.8) >was it fun?<
6           (0.9)
7     A:    it wa:s okay we went and ate subway °a:nd° (1.2)
8           yeah just chatted about (0.2) world events
9           and [the   economies    ]
10    B:        [just >chilling out<]
11          (0.3)
12    A:    yeah
13          (1.0)
14    B:    cool (0.2) yeah I ah (1.0) >what did I do I just<
15          studied (0.2) >spent the whole weekend studying
16          did semantics on Saturday, (0.6) and di:d (0.2) CA
17          >conversation analysis< on ↑Sun↓day
18          (1.8)
19    A:    sounds like fun °there goes° the students life
20          HA ha [.hh hh .hh hh  .hh  ]
21    B:          [ah:: yes it was very] interesting
22          (0.8)
23    A:    um ↑hm
24          (3.3)
25    A:    °so yeah°
```

In the same way as we saw in example (2), speaker A's inquiry about what speaker B has been doing (line 1) elicits a telling about his weekend (lines 3-10), but the inquiry is not reciprocated by speaker B. After a gap of silence in line 13, speaker A launches his own telling through a self-directed inquiry (line 14: what did I do), before going on to talk about his own weekend's activities. The students labelled this interactional practice "occasioning of self-talk". The practice involves the speaker directing a question to the recipient, which subsequently furnishes grounds for the speaker to "tell an experience" (cf. Pomerantz, 1980, who describes the practice of "telling my side" as a fishing device, which involves the speaker "telling an experience" as a "possible elicitor of information" from the recipient, p. 187).

The practice of ironically receipting complainables was first noted by the student (an L2 speaker) in her own recording. The practice involves speaker A responding to a potential complaint from speaker B with ironic uptake. This occasions recognition on the part of speaker B of both the complainable import of his/her prior utterance, as well as the move by speaker A to a non-serious, ironic frame. This recognition is displayed by speaker B responding with a further ironic utterance subsequent to speaker A's initial ironic formulation. The student then found another example of the same practice in the excerpt from GCSAusE09 reproduced above. In lines 15-17, speaker B describes how he spent the whole weekend studying. This is treated as a complainable by speaker A, who responds with an ironic formulation in line 19 ("sounds like fun °there goes° the students life"). It is recognisably ironic as clearly speaker A does not mean studying all weekend is a fun thing to be doing, but rather that it is exactly the opposite, and thus something about which making a complaint is reasonable. It is also marked as being delivered within a non-serious frame as speaker A also initiates laughter (line 20). However, instead of reciprocating the laughter, speaker B responds with another ironic formulation in line 21 ("ah:: yes it was very interesting"), where he displays support for speaker A's previous stance.

According to the results of a brief written feedback survey about the corpus, which was distributed at the end of the course to the 24 students who were enrolled (see Appendix A), approximately 90% of the students accessed the GCSAusE online, with 82% of them going on to use data from the corpus in their analytical projects. While 82% of the students reported the corpus was easy to use, most of the students

accessed data in the corpus through the browse function, with only 40% using a keyword or guided search.

In order to get more detailed feedback an in-depth focus group was conducted with a smaller number of students. Three students took part in the focus group, two L1 speakers and one L2 speaker, with the second author facilitating the discussion. The second author was not involved in teaching or grading the course in order to minimize any possible conflicts of interest in conducting this evaluative focus group (see Appendix B for a list of the guiding questions). The discussion was recorded and transcribed, and then analysed independently and the findings subsequently compared by the two authors. Three key themes emerged from this analysis. First, the students emphasised the importance of having access to the original audio recordings, not just the transcripts. One L1 student, for instance, in response to being asked how she used the corpus said:

(4) I went into ones that had audio because I find it really hard to read the transcripts and just try and read it so I went through and found ones that had, so I just browsed and found ones that had audio and then I went through so yeah that's how I went through.

While CA transcripts are useful for detailed analysis, they can be challenging for students to interpret, particularly if no accompanying audio file is available, a point that was also reiterated by the L2 student.

A second theme that emerged was how getting access to other conversations, beyond the one they themselves recorded and transcribed, helped the students to not only appreciate the complex nature of conversational interaction, but also its ability to stir interest in learning more about it. In the following excerpt, an L1 student responded to the facilitator's question about the benefits of donating to or using the corpus (turn 24) by claiming she found examining other conversations peaked her interest in analysing conversational structure (turn 25).

(5)

24	A:	And what benefits can you see for you from donating or using the corpus?
25	B:	I think well immediately for other research projects but I guess it's really interesting to read them, like you go through and the little subtleties in the conversation I guess, I don't know it kind of opens up your eyes to what really happens in a conversation because even though I didn't think mine was that interesting, then when I kind of looked at others, I was like wow mine is kind of, it's really, I don't know, it kind of opens you up to what conversations are about, and I guess …
26	D:	Yeah at first you don't think it is interesting but then …
27	B:	Yeah.
28	D:	But when you look in more details in every conversation you can find something.

This claim was supported by the L2 student, who said she did not find analysing the conversations interesting at first (turn 26), but later appeared to find some value in doing so (turn 28).

The third theme was that the students found the corpus useful in undertaking their projects, as they were given access to a greater amount of data than they could have feasibly gathered on their own. However, the students also recognised that the corpus could be used by the wider community, suggesting a sense of belonging to the university research community engendered through contributing to the corpus.

(6)

32 C: I think it's just the very practical way to like show people's work and data even if it's not just for assessment, like it's just sort of like the whole bringing together of like all the students and other people's research, I think that is very practical and it's like all in one area as well. Rather than having to go and search and find say for example we need to get Australian English examples to try and search for that, like it's all sort of basically there for you, it's just all really practical.

33 B: I think it's good and like anyone that, even having it there and if you have another interest, if you take the course and think oh that's what I really want to do, you can kind of go back there and […] it's good to have like the opportunity for Griffith students and then like the wider community to have them being able to research and use our stuff.

34 A: Okay

35 C: Sort of puts Griffith on show as well I think what's actually happening, so.

As can be seen in the excerpt above, one student mentions the practicality of the corpus (turn 32), while another points to the potential for wider use of the corpus by others outside of the university (turns 33 and 35). It appears that in having the opportunity to be both creators and users of the corpus, and see how it results in the real and ongoing accumulation of knowledge about language use, students gained a sense of having a place within the research community.

6. Concluding remarks

The Griffith Corpus of Spoken Australian English (GCSAusE) has been, and continues to be, created through a cyclical and collaborative model. In this way, students are both creators and users of the corpus. This approach is one possible means of overcoming the current bottleneck in readily accessing authentic spoken interaction for use in instructional pragmatics. It also forms part of the research-based learning paradigm implemented in this and other related courses in the International English program, where undergraduate students, both L1 and L2 speakers of English, have the opportunity to have their research projects published online.[9] While the corpus itself is created and used in an advanced level of course in which both L1 and L2 students take part, L2 students in other English courses are given access to the corpus as well. It is this latter use of the GCSAusE which means that such a model can benefit not only educational contexts where research-based learning is feasible, but also educational settings where access to "real life" interactional data is limited, since it contributes in a very real way to progressively increasing the amount of authentic interactional data available for use in ESL/EFL classrooms.

There are, of course, likely to be some limitations to the implementation of this kind of model, but we believe its basic principles can be at least partially adapted to other contexts. For instance, if students are at a lower level of proficiency in the L2

in question, they can be guided to produce less complex transcriptions. If access to recording authentic interactions in the target language is not readily available, Krafts and Geluykens (2008) propose that examples of spontaneous interaction can be found in new television formats such as “docusoaps” or “fly-on-the-wall documentaries”, which attempt to “simply record everyday events as they unfold, without any script or manipulation” (p. 101). The latter claim can be disputed perhaps, but it is clear that the Internet increasingly provides access to all kinds of spontaneous interactions across different languages. The model proposed here allows teachers and learners alike to systematically exploit this potential, whilst also building an ongoing resource for the institution where the learning takes place. In this way, students are able to learn through sharing the fruits of their efforts with others.

References

Adolphs, S. (2008). *Corpus and Context. Investigating Pragmatic Functions in Spoken Discourse*. Amsterdam: John Benjamins.

Adolphs, S., & Carter, R. (2007). Beyond the word. New challenges in analysing corpora of spoken English. *Eourpean Journal of English Studies, 11*(2), 133-146.

Allwood, J. (2008). Multimodal corpora. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics. An International Handbook. Volume 1* (pp. 207-225). Berlin: Mouton de Gruyter.

Archer, D., Culpeper, J., & Davies, M. (2008). Pragmatic annotation. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics. An International Handbook. Volume 1* (pp. 613-642). Berlin: Mouton de Gruyter.

Arundale, R. (1999). An alternative model and ideology of communication for an alternative to politeness theory. *Pragmatics, 9*(1), 119-154.

Arundale, R. (2005). Pragmatics, conversational implicature, and conversation. In K. Fitch & R. Sanders (Eds.), *Handbook of Language and Social Interaction* (pp. 41-63). Mahwah, NJ: Lawrence Erlbaum.

Arundale, R. (2010). Constituting face in conversation: face, facework and interactional achievement. *Journal of Pragmatics, 42*(8), 2078-2105.

Bardovi-Harlig, K. (2006). On the Role of Formulas in the Acquisition of L2 Pragmatics. In K. Bardovi-Harlig, J. Félix-Brasdefer & A. Omar (Eds.), *Pragmatics Language Learning. Volume 11* (pp. 1-28). Honolulu: National Foreign Langauge Resource Centre, University of Hawai’i at Mānoa

Bardovi-Harlig, K. (2009). Conventional expressions as a pragmalinguistic resource: recognition and production of conventional expressions in L2 pragmatics. *Language Learning, 59*(4), 755-795.

Bardovi-Harlig, K. (2010). Recognition of conventional expressions in L2 pragmatics. In G. Kasper, H. Nguyen, D. Yoshimi & J. Yoshioka (Eds.), *Pragmatics Language Learning. Volume 12* (pp. 141-162). Honolulu: National Foreign Language Resource Center, University of Hawai’i at Mānoa.

Barraja-Rohan, A.-M. (1997). Teaching conversation and sociocultural norms with conversation anaylsis. *Australian Review of Applied Linguistics Series, 14*, 71-88.

Brinckmann, C. (2009). Transcription bottleneck of speech corpus exploitation. In V. Lyding (Ed.), *Proceedings of the Second Colloquium on Lesser Used Languages and Computer Linguistics (LULCLII 2008)* (pp.165-179). Bolzano/Bozen: EURAC.

Čermák, F. (2009). Spoken corpora design. Their constitutive parameters. *International Journal of Corpus Linguistics, 14*(1), 113-123.
Chambers, A. (2007). Integrating Corpora in Language Learning and Teaching. *ReCall, 19*(3), 249-251.
Clancy, B. (2011). *Do you want to do it yourself like?* Hedging in Irish traveller and settled family discourse. In B. Davies, M. Haugh & A. Merrison (Eds.), *Situated Politeness* (pp. 129-146). London: Continuum.
Crowdy, S. (1993). Spoken corpus design. *Literary and Linguistic Computing, 8*(4), 259-265.
Crowdy, S. (1994). Spoken Corpus Transcription. *Literary and Linguistics Computing, 9*(1), 25-28.
Culpeper, J. (2011). *Impoliteness: Using Language to Cause Offence*. Cambridge: Cambridge University Press.
Davies, M. (2009). The 385+ million word *Corpus of Contemporary American English* (1990-2008+). *International Journal of Corpus Linguistics, 14*(2), 159-190.
Deppermann, A., & Schütte, W. (2008). Data and transcription. In G. Antos & E. Ventola (Eds.), *Handbook of Interpersonal Communication* (pp. 179-213). Berlin: Mouton de Gruyter.
Félix-Brasdefer, J. (2006). Using the negotiation of multi-turn speech acts: using conversation-analytic tools to teach pragmatics in the FL classroom. In K. Bardovi-Harlig, J. Félix-Brasdefer & A. Omar (Eds.), *Pragmatics and Language Learning. Volume 11* (pp. 165-198). Honolulu, Hawai'i: National Foreign Language Resource Center, University of Hawai'i at Manoa.
Félix-Brasdefer, J. (2008). *Politeness in Mexico and the United States: A Contrastive Study of the Realization and Perception of Refusals.* Amsterdam: John Benjamins.
Geluykens, R., & Kraft, B. (2008). The use(fulness) of corpus research in cross-cultural prgamatics: Complaining in intercultural service encounters In J. Romero-Trillo (Ed.), *Pragmatics and Corpus Linguistics* (pp. 93-117). Berlin: Mouton de Gruyter.
Haugh, M. (2009). Designing and multimodal spoken component of the Australian National Corpus. In M. Haugh, K. Burridge, J. Mulder & P. Peters (Eds.), *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus: Mustering Languages* (pp. 74-86). Sommerville, MA: Cascadilla Press.
Haugh, M. (2010). Co-constructing what is said in interaction. In E. Nemeth T & K. Bibok (Eds.), *The Role of Data at the Semantics-Pragmatics Interface* (pp. 349-380). Berlin: Mouton de Gruyter.
Haugh, M. (2012). Conversational interaction. In K. Allan & K. Jaszczolt (Eds.), *Cambridge Handbook of Pragmatics* (pp. 251-273). Cambridge: Cambridge University Press.
Howe, J. (2006). The rise of crowdsourcing. *Wired, 14*(6), 1-5. Available at http://www.wired.com/wired/archive/14.06/crowds.html.
Huth, T., & Taleghani-Nikazm, C. (2006). How can insights from conversation analysis be directly applied to teaching L2 pragmatics? *Language Teaching Research 10*(1), 53-79.
Ishihara, N. (2010). Instructional pragmatics: bridging teaching, research, and teacher education. *Language and Linguistics Compass, 4*(10), 938-953.

Ishihara, N., & Cohen, A. (2010). *Teaching and Learning Pragmatics: Where Language and Culture Meet*. Harlow, UK: Longman.

Jefferson, G. (2004). Glossary of transcript symbols with an introduction. In G. Lerner (Ed.), *Conversation Analysis: Studies from the First Generation* (pp. 13-23). Amsterdam: John Benjamins.

Jeon, E. H., & Kaya, T. (2006). Effects of L2 instruction on interlanguage pragmatic development: a meta-analysis. In J. M. Norris & L. Ortega (Eds.), *Synthesizing Research on Language Learning and Teaching* (pp. 165-211). Amsterdam: John Benjamins.

Jiang, X. (2006). Suggestions: what should ESL students know? *System, 34*(1), 36-54.

Jucker, A. H., Schreier, D., & Hundt, M. (Eds.). (2009). *Corpora: Pragmatics and Discourse*. Amsterdam: Rodophi.

Kasper, G. (2001). Four perspectives on L2 pragmatic development. *Applied Linguistics, 22*(4), 502-530.

Kasper, G. (2006). Speech acts in interaction: towards discursive pragmatics. In K. Bardovi-Harlig, C. Felix-Brasdefer & A. S. Omar (Eds.), *Pragmatics and Language Learning Volume 11* (pp. 281-314). Honolulu: National Foreign Language Resource Center, University of Hawai'i at Manoa.

Kasper, G., Nguyen, H., & Yoshimi, D. (2010). Introduction. In G. Kasper, H. Nguyen, Y. D. Rudolph & J. Yoshioka (Eds.), *Pragmatics Language Learning. Volume 12* (pp. 1-14). Honolulu: National Foreign Language Center, University of Hawai'i at Mānoa.

Kasper, G., & Rose, K. (2002). *Pragmatic Development in a Second Language*. Malden, MA: Blackwell.

Koester, A. (2002). The performance of speech acts in workplace conversations and the teaching of communicative functions. *System, 30*, 167-184.

Koester, A. (2009). Conversation analysis in the language classroom. In S. Hunston & D. Oakey (Eds.), *Introducing Applied Linguistics: Key Concepts and Skills* (pp. 37-48). London: Routledge.

Marra, M. (2008). Recording and analysing talk across cultures. In H. Spencer-Oatey (Ed.), *Culturally Speaking. Culture, Communication and Politeness Theory* (pp. 304-321). London: Continuum.

Newton, J. (2004). Face-threatening talk on the factory floor: Using authentic workplace interactions in language teaching. *Prospect, 19*(1), 47-64.

O'Keefe, A., McCarthy, M., & Carter, R. (2007). *From Corpus to Classroom. Language Use and Language Teaching*. Cambridge: Cambridge University Press.

Peters, P. (2009). The architecture of a multipurpose Australian National Corpus. In M. Haugh, K. Burridge, J. Mulder & P. Peters (Eds.), *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus: Mustering Languages* (pp. 1-9). Sommerville, MA: Cascadilla Press.

Pomerantz, A. (1980). Telling my side: "limited access" as a "fishing" device. *Sociological Inquiry, 50*, 186-198.

Ramírez-Verdugo, M. D. (2008). A cross-linguistic study on the pragmatics of intonation *in* directives. In J. Romero-Trillo (Ed.), *Pragmatics and Corpus Linguistics* (pp. 205-233). Berlin: Mouton de Gruyter.

Rohlfing, K., Loehr, D., Duncan, S., Brown, A., Franklin, A., Kimbara, I., et al. (2006). Comparison of multimodal annotation tools - workshop report. *Gesprachsforschung - Online-Zeitschrift zur verbalen Interaktion, 7*, 99-123.

Romero-Trillo, J. (Ed.). (2008). *Pragmatics and Corpus Linguistics.* Berlin: Mouton de Gruyter.
Rose, K. (2005). On the effects of instruction in second language pragmatics. *System, 33*, 385-399.
Rose, K., & Ng, C. (2001). Inductive and deductive teaching of compliments and compliment responses. In K. Rose & G. Kasper (Eds.), *Pragmatics in Language Teaching* (pp. 145-170). Cambridge: Cambridge University Press.
Rühlemann, C. (2010). What can a corpus tell us about pragmatics? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 288-301). London: Routledge.
Schauer, G. A., & Adolphs, S. (2006). Expressions of gratitude in corpus and DCT data: Vocabulary, formulaic sequences, and pedagogy. *System, 34*, 119-134.
Schmidt, T. (2004). *Transcribing and annotating spoken language with EXMARaLDA.* Paper presented at the LREC Workshop on XML based richly annotated corpora, Lisbon.
Schmidt, T., & Schütte, W. (2010). FOLKER: an annotation tool for efficient transcription of natural, multi-party interaction *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)* (pp. 2091-2096). Valletta, Malta: European Language Resources Association (ELRA).
Takimoto, M. (2008). The effects of deductive and inductive instruction on the development of language learners' pragmatic competence. *The Modern Language Journal, 92*(3), 369-386.
Taleghani-Nikazm, C. (2002). A conversation analytical study of telephone conversation openings between native and nonnative speakers. *Journal of Pragmatics, 34*, 1807-1832.
Taylor, C. (2009). Interacting with conflicting goals. Facework and impoliteness in hostile cross-examination. In J. Morely & P. Bayley (Eds.), *Corpus Assisted Discourse Studies on the Iraq Conflict: Wordings of War* (pp. 208-233). London: Routledge.
Taylor, C. (2011). Negative politeness forms and impoliteness functions in institutional discourse: a corpus-assisted approach. In B. Davies, M. Haugh & A. Merrison (Eds.), *Situated Politeness* (pp. 209-231). London: Continuum.
Thompson, P. (2004). Spoken language corpora. In M. Wynne (Ed.), *Developing Linguistic Corpora: a Guide to Good Practice* (pp. 59-70). Oxford: Oxbow Books.
Usami, M. (2005). Why do we need to analyse authentic materials in developing conversation teaching materials? In Y. Kawaguchi, S. Zaima, T. Takagi, K. Shibano & M. Usami (Eds.), *Linguistic Informatics. State of the Art and the Future* (pp. 279-294). Amersterdam: John Benjamins.
Vaughan, E. (2008). 'Got a date or something?': A corpus analysis of the role of humour and laughter in the workplace meetings of English language teachers. In A. Ädel & R. Reppen (Eds.), *Corpora and Discourse: The Challenge of Different Settings* (pp. 95-115). Amsterdam: John Benjamins.
Vine, B. (2004). *Getting Things Done at Work.* Amsterdam: John Benjamins.
Voormann, H., & Gut, U. (2008). Agile corpus creation. *Corpus Linguistics and Linguistic Theory, 4*(2), 235-251.
Wichmann, A. (2008). Speech corpora and spoken corpora. In A. Ludeling & M. Kytö (Eds.), *Corpus Linguistics. An International Handbook. Volume 1* (pp. 187-207). Berlin: Mouton de Gruyter.

Wong, J. (2002). “Applying” conversation analysis in applied linguistics: Evaluating dialogue in English as a second language textbooks. *IRAL, International Review of Applied Linguistics in Language Teaching, 40*(1), 37-60.
Wong, J., & Waring, H. Z. (2010). *Conversation Analysis and Second Language Pedagogy*. London: Routledge.
Wulff, S. (2010). *Rethinking Idiomaticity. A Usage-based Approach*. London: Continuum.
Yates, L. (2010). Dinkas Down Under: Request Performance in Simulated Workplace Interaction. In G. Kasper, H. Nguyen, D. Yoshimi & J. Yoshioka (Eds.), *Pragmatics Language Learning. Volume 12* (pp. 113-140). Honolulu: National Foreign Language Resource Center, University of Hawai’i at Mānoa.

Appendix A: Survey of Griffith Corpus of Spoken Australian English (GCSAusE)

1. Did you access the GCSAusE online?
   Yes / No

2. Did you use data from the GCSAusE in your research project?
   Yes / No

3. Did you find the corpus easy to use?
   Yes / No
Why or why not?

4. How did you use the corpus?
   a. Browse: Yes / No
   b. Keyword search: Yes / No
   c. Guided search: Yes / No

5. How do you think the corpus could be improved?

Appendix B: Focus group guiding questions

1. How are you using data from the corpus in your research project?

2. How did you find suitable data in the corpus (e.g. browse, guided search etc.).

3. Did you find it easy to locate suitable data? Why/why not?

4. Did you donate data to the corpus? Why/why not?

5. What benefits can you see for you from donating or using the corpus?

6. How do you think the corpus could be improved?

[1] cf. Kasper, Nguyen and Yoshimi (2010) who limit the scope of pragmatics to the “study of language-mediated social *action*” (p. 3, emphasis added), reflecting a more strictly CA-oriented approach to pragmatics.

[2] Such limitations account for why Barraja-Rohan (1997) and Félix-Brasdefer (2006) suggest drawing from mainstream pragmatics as well as CA in the teaching of L2 pragmatics. However, such an approach is potentially fraught with problems, namely, that drawing conclusions using rationalistic theories of pragmatics within the context of a constructivist approach to data analysis arguably generates theoretical and methodological incoherence (Arundale, 2005, 2010; Haugh, 2010, pp. 372-3).

[3] See: http://www.sscnet.ucla.edu/soc/faculty/schegloff/sound-clips.html, and http://talkbank.org/CABank/.

[4] See: http://www.linguistics.ucsb.edu/research/sbcorpus.html

[5] Further information about the GCSAusE, including detailed metadata can be found by logging in as a guest at http://equella.rcs.griffith.edu.au/research/logon.do. The corpus is also going to be made available to researchers and educators more widely through the Australian National Corpus, which is currently being established in a joint venture between Griffith University and Macquarie University with funding from the Australian National Data Service. For further information see http://www.ausnc.org.au.

[6] To blur the face of the participants would defeat one of the main purposes of having audiovisual recordings in the first place, namely, to allow the analysis of gaze, facial expressions and so on. While the physical setting and posture and some gestures of the participants could be retained in this way, the extra effort involved was not considered worthwhile if only such restricted data could be made available.

[7] A new type of specialised transcription software, FOLKER (Schmidt and Schütte, 2010), has since been released, however, which may prove more accessible to students. See http://agd.ids-mannheim.de/html/folker_en.shtml for further details.

[8] Audacity, a freely available sound editor is available for download from http://audacity.sourceforge.net/.

[9] See the online journal, *Griffith Working Papers in Pragmatics and Intercultural Communication*, available at http://www.griffith.edu.au/arts-languages-criminology/school-languages-linguistics/publications.